National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

FRONT OFFICE FOOD AND PRODUCT SAFETY

Assessment of the method of Stacked Model Averaging from the point of view of its application in determining minimal eliciting doses of food allergens

Assessment requested by:	Jacqueline Castenmiller (Office for Risk Assessment & Research (BuRO))
Assessment performed by:	RIVM
Date of request:	23-02-2022
Date of assessment:	06-04-2022 (draft) 04-05-2022 (final)
Project number:	V/093130

Subject

In 2016, the Office for Risk Assessment & Research (BuRO) of the Netherlands Food and Consumer Product Safety (NVWA) published an advice on preliminary reference doses for allergens in foods (BuRO, 2016). This was based on the lowest eliciting doses (EDs) that were derived from published information from the TNO/FARRP database (Taylor et al., 2014; Remington, 2013).

In 2020, the EDs have been recalculated by Remington et al. using data from the extended TNO/FARRP database (Remington et al., 2020). A newly developed statistical method Stacked Model Averaging (Wheeler et al., 2019) was used by Remington et al. (2020) to derive these EDs.

Questions

1. Is the method of Stacked Model Averaging described by Wheeler et al. (2019) suitable for estimating minimal EDs?
2. Does the application of this method result in an over- or underestimation of minimal EDs?
3. Are there points for attention when using the model and applying its results?

Conclusions

1) The method is not unsuitable to estimate minimal EDs. But it is not preferable to the methods used earlier (based on fitting various distributions to the data and obtaining estimates of doses from them) because in order to check the outcome of the new method one needs in essence to use earlier methods.

2) The method does not underestimate or overestimate derived minimal EDs compared to the methods used earlier, provided that the fit of the several component distributions be checked in order to help judging the derived minimal EDs, as is done with earlier methods.

3) When using the method, it is recommended that the fit of the several component distributions be checked in order to help judging the derived minimal EDs, as is done with earlier methods.

Introduction

In 2016, the Office for Risk Assessment & Research (BuRO) of the Netherlands Food and Consumer Product Safety (NVWA) published an advice on preliminary reference doses for allergens in foods (BuRO, 2016). This advice was based on the lowest eliciting doses (EDs) that were available at that time. These EDs were based on published information from the TNO/FARRP database (Taylor et al., 2014; Remington, 2013). The EDs were derived based on data on no-observed-adverse-effect levels (NOAELs) and lowest-observed-adverse-effect levels (LOAELs) from clinical challenges of allergic individuals. These EDs were derived by fitting three distributions (log-normal, log-logistic and Weibull) to the data. BuRO used the lowest predictive ED01 or the lower 95% confidence limit of the ED05 of one of the three models (BuRO, 2016).

In 2020, the minimal EDs have been recalculated by Remington et al. with data from the extended TNO/FARRP database (Remington et al., 2020). Here, a newly developed statistical method was used for interval-censored data. This method, Stacked Model Averaging, is a Bayesian approach for the analysis of interval-censored data. In it, multiple parametric survival estimates are combined by means of weighted averages into a single survival estimate (Wheeler et al., 2019).

Question 1: Is the method of Stacked Model Averaging described by Wheeler et al. (2019) suitable for estimating minimal EDs?

The method of Stacked Model Averaging is not unsuitable to estimate minimal EDs. But we do not recommend that it replace the methods used so far, because, in order to check the outcome of the new method, one needs in essence to use the methods used until now. The reasons for this are elaborated in the answer to Question 3.

In the methods used so far, a number of distributions are fitted to NOAEL and LOAEL data, which provide estimates of doses (each distribution provides one estimate)[1]. The minimal ED is selected from these estimates by taking the goodness-of-fit of the distributions into account together with expert knowledge and intuition (from whoever applies the method and/or decides which estimates of minimal EDs to adopt).

In Stacked Model Averaging too one fits a set of distributions. But in addition an average distribution is derived considering the relative validity of the individual distributions. Distributions that fit or predict the data better have more weight in the average distribution. An estimate of the ED is then obtained from that average distribution. So the essential difference between the Stacked Model Averaging approach and the earlier methods is that the former replaces a set of distributions and associated dose estimate by a single distribution and the single associated dose estimate.

Question 2: Does the application of this model result in an over- or underestimation of minimal EDs?

[1] For the definition of NOAEL and LOAEL data see, for example, the introduction to Taylor *et al.* (2009).

Relative to the doses derived by earlier methods, there seems to be little risk of the minimal EDs being overestimated by Stacked Model Averaging, provided the plausibility of the component distributions involved in it be carefully checked. Earlier methods rely on expert judgement based on the goodness of fit of the distributions, because the only way to gauge the plausibility of the various distributions is to check how well they fit the data. So the goodness of fit of the distributions is the basis for decision-making. Even though in Stacked Model Averaging a single distribution and therefore a single dose estimate is derived, it is still of great importance to identify and examine the fitted distributions with greater (and lesser) weight in the final estimate, in order to check the reliability of the results. This requires the same kind of expert judgement involved in earlier methods. Given the similarity of the very fundamental aspects of the two approaches (namely the fitting different distributions and the need for expert judgement), no under- or overestimation is expected regarding the minimal EDs derived by Stacked Model Averaging. Still, we should emphasize the importance of examining —and exhibiting clearly to the targeted audience — the goodness of fit of the various distributions involved and the arguments for adopting estimates of minimal EDs from them. In order to see why this is important, we give some concrete examples.

In subsection 5.1 of Wheeler et al. (2019, 2021), the new estimates of ED01 for peanuts are basically the same as the old ones, and if the new estimates of ED05 and ED10 are about twice as large as the old ones then there ought to be good grounds for discounting the old ones due to a poorer model fit. However, in this case we have no means of checking and comparing the fits of the various distributions that result from the application of the new method in Wheeler et al. (2019, 2021), because the graphs of figures 1 and 2 are not as clear and detailed as those shown in Remington (2013), say as figure A on p. 107, because we do not possess the data and hence have not looked at them ourselves, and because we cannot assess the relative credibility of the various data sets and of the studies those data sets come from. Based on the figure (i.e. figure A just mentioned), and taking the data in it for granted, our *impression* is that the log-logistic model and its estimate of 0.13 for the value of ED01 are slightly closer to the truth than the log-normal and corresponding estimate of 0.28 for the value of ED01. It is probable, judging from the sample size and the graph, that both estimates of 0.13 and 0.28 are consistent with the data, so that it is not possible to decide between them without further data; in this case, proposing the smaller, 0.13, would seem safer to us, but more we cannot say. Still in connection with this example, in subsection 5.1 of Wheeler et al. (2019, 2021) it is said that "Unsurprisingly, the stacked survival estimates, given in figure 2, are close to the previous accelerated failure time results"; but we cannot see that this is the case nor which of the distributions, if any, used earlier by Remington (2013)—Weibull, log-normal and log-logistic—is closer to the stacked survival estimate. If in the application of the new method of Wheeler et al. (2019, 2021) the weights given to the Weibull, generalized Pareto and log-Laplace add up practically to 1 (0.83, 0.11 and 0.06, though the text states "0.83, 0.11, and 0.06%") then the weight given to the log-normal (log-Gaussian) must be close to zero, which may be surprising in view of the fits obtained earlier by Remington (2013) and shown in the figure A we have been mentioning; but we cannot be certain about how surprising or contradictory this is because, among other things, we are not sure that the data used by Wheeler et al. (2019, 2021) are *exactly* the same as the data used earlier by Remington (2013).

Thus it appears from this example that in applications of Stacked Model Averaging the results may not be completely clear. For another example, we may note that on p. 8 of Remington et al. (2020) it is argued that the new estimate of ED01 for egg is considerably larger than the earlier one: in this case, it is seen from figure A on p. 114 of Remington (2013) that the Weibull distribution, which in the earlier method gave the lowest estimates, fits the lower range of the data better than do the other two distributions, so that, if used with the same data as before (as implied by the discussion on p. 8 of Remington et al., 2020, though apparently contradicted by the statement that "all three statistical models

[…] fit the egg data quite well"), the new method would seem to give too little weight to the Weibull distribution.

All in all, however, and provided the goodness-of-fit of the various distributions be made transparent and their plausibility assessed, it seems that uncertainties and concerns about bias in estimates should have less to do with the statistical method than with the representativeness of the populations sampled and with the validity of the measurements, both of which are brought into question in section 2 of Taylor et al. (2009), for example.

Question 3: Are there points for attention when using the model and applying its results?

When using the method, it is recommended that the fit of the several component distributions be checked in order to help judging the derived minimal EDs as is done in the methods used so far. This is because of the added complexity and the reduced transparency of the Stacked Model Averaging approach, as explained next.

One supposed advantage of the Stacked Model Averaging approach is the generation of a single weighted averaged distribution and therefore a single estimate of the minimal ED. This, however, is achieved by building quite some structure on top of the selected distributions and making additional assumptions. Such assumptions may not be realistic and are not verifiable. Meanwhile by focusing on the single weighted distribution produced by the Stacked Model Averaging method assessors may ignore the uncertainties made obvious by earlier methods, reducing the transparency of the decision process.

Conclusions

1) The method is not unsuitable to estimate minimal EDs. But it is not preferable to the methods used earlier because in order to check the outcome of the new method one needs in essence to use the earlier methods.

2) The method does not underestimate or overestimate derived minimal EDs compared to the methods used so far, provided that the fit of the several component distributions be checked in order to help judging the derived minimal EDs, as is done in earlier methods.

3) When using the method, it is recommended that the fit of the several component distributions be checked in order to help judging the derived minimal EDs, as is done in the methods used so far.

References

BuRO (2016). Advice of Buro on preliminary reference doses for food allergens. Available via: https://english.nvwa.nl/documents/consumers/food/safety/documents/advice-of-buro-on-preliminary-reference-doses-for-food-allergens

Remington, B.C. (2013). Risk Assessment of Trace and Undeclared Allergens in Processed Foods. Dissertations, Theses & Student Research in Food Science and Technology, 32. Available at https://digitalcommons.unl.edu/foodscidiss/32.

Remington, B.C., Westerhout, J., Meimaa, M.Y., Bloma, W.M., Kruizinga, A.G., Wheeler, M.W., Taylor, S.L., Houbena, G.F. and Baumert, J.L. (2020). Updated population minimal eliciting dose distributions for use in risk assessment of 14 priority food allergens. Food and Chemical Toxicology, 139: 111259.

Taylor, S.L., Baumert, J.L., Kruizinga, A.G., Remington, B.C., Crevel, R.W.R., Brooke-Taylor, S., Allen, K.J., The Allergen Bureau of Australia & New Zealand, Houben, G. (2014). Establishment of Reference Doses for residues of allergenic foods: Report of the VITAL Expert Panel. Food and Chemical Toxicology, 63: 9-17.

Taylor, S.L., Crevel, R.W.R., Sheffield, D., Kabourek, J. and Baumert, J. (2009). Threshold dose for peanut: Risk characterization based upon published results from challenges of peanut-allergic individuals. Food and Chemical Toxicology, 47: 1198–1204.

Wheeler, M.W., Westerhout, J., Baumert, J.L. and Remington, B.C. (2019). Bayesian stacked parametric survival with frailty components and interval censored failure times. Working paper posted at arXiv, available at https://arxiv.org/abs/1908.11334.